MATH5633 Loss Models I Autumn 2024

Chapter 1: Basic Probability

Lecturer: Kenneth Ng

Preview

MATH 5633 Loss Model I aims to equip students with the mathematical foundations for the Exam FAM-S offered by the Society of Actuaries (SOA). As suggested by the name of the examination, this course will focus on the construction of actuarial models associated with short-term insurance products, such as health, property and casualty (P&C), group insurance, and travel insurance etc. In this chapter, we shall go over the basic probability tools that are essential for this course.

Key topics in this chapter:

1. Probability measures and random variables;
2. Distributional quantities, such as raw moments, central moments, and percentiles;
3. Moment generating functions and probability generating functions;
4. Conditional distributions and conditional expectations

1 Probability

We consider a set Ω, called the *sample space*, which contains all the possible outcomes of a random experiment. A probability is a measure of likelihood that a specific event will occur.

Definition 1.1 A ***probability measure*** $\mathbb{P}$ is a function defined on subsets of the sample space Ω such that the following properties are satisfied:

1. for any $A \subset \Omega$, $0 \leq \mathbb{P}(A) \leq 1$;
2. $\mathbb{P}(\varnothing) = 0$ and $\mathbb{P}(\Omega) = 1$;
3. for any $A, B \subset \Omega$ with $A \cap B = \varnothing$, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

The following useful properties can be deduced from the definition of probability, whose proof are left as an exercise.

1. (Inclusion-Exclusion) For any $A, B \subset \Omega$, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

2. (Complementary Event) For the *complementary event* A^c of A, where $A^c := \Omega \backslash A$, it holds that $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.

3. (Monotonicity) For any sets $A \subset B$, $\mathbb{P}(A) \leq \mathbb{P}(B)$.

4. (Law of Total Probability) Let $\{B_i\}_{i=1}^n$ be a collection of *mutually exclusive* and *exhaustive* events, i.e., $B_i \cap B_j = \varnothing$ for $i \neq j$, and $\cup_{i=1}^n B_i = \Omega$. Then, for any $A \subset \Omega$, $\mathbb{P}(A) = \sum_{i=1}^n \mathbb{P}(A \cap B_i)$.

Remark 1.1. *(Impossible versus zero probability)* An event $A \subset \Omega$ with zero probability (i.e., $\mathbb{P}(A) = 0$) does NOT mean A is impossible to happen. For example, consider a random experiment of picking a number from the interval $\Omega = [0, 1]$ uniformly. The probability of picking any given number would be zero. However, you will end up picking a specific number, despite the probability of getting it is zero. This is because the sample space Ω contains a continuum of possible outcomes. Similarly, $\mathbb{P}(A) = 1$ does not mean A must occur. In measure theory, we say that A occurs *almost surely*.

Example 1.1 A survey indicates that 60% of citizens in Westeros have purchased health insurance products, 45% of them have purchased variable annuity products, and all of them have purchased at least one of the above two products. Find the proportion of citizens in Westeros who have purchased a variable annuity but not a health insurance.

Solution:

Let A be the event of purchasing health insurance products, and B be the event of purchase a variable annuity. We know that $\mathbb{P}(A) = 0.6$, $\mathbb{P}(B) = 0.45$, and $\mathbb{P}(A \cup B) = 1$. We are to find $\mathbb{P}(A^c \cap B)$. By inclusion-exclusion,

$$1 = \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) = 0.6 + 0.45 - \mathbb{P}(A \cap B),$$

which gives $\mathbb{P}(A \cap B) = 0.05$. Hence,

$$\mathbb{P}(A^c \cap B) = \mathbb{P}(A) - \mathbb{P}(A \cap B) = 0.6 - 0.05 = 0.55.$$

2 Conditional Probability

Conditional probability describes the likelihood of occurrence of an event A given the occurrence of another event B. When the two events A and B are related, the occurrence of B would change the probability of the occurrence of A. For instance, the probability that a 25-year old male who has lung cancer could be as low as 2%. However, if we know that the male is a smoker, the probability could be doubled.

Definition 2.1 Suppose that $A, B \subset \Omega$ with $\mathbb{P}(B) > 0$. Then, the **conditional probability** of A given B is defined as

$$\boxed{\mathbb{P}(A|B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.}$$

Given that B has occurred, we need to confine ourselves to the smaller population set B, and searches for the outcome in A therein. Hence, in the definition, we are using $\mathbb{P}(B)$ as a numeraire in the denominator, and $\mathbb{P}(A \cap B)$ as the occurrence probability within B.

If the occurrence of B does not change the probability of A, then A and B are said to be *independent*.

Definition 2.2 The events A and B are said to be ***independent*** if

$$\boxed{\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).}$$

Remark 2.1.

1. If $\mathbb{P}(B) > 0$, then A and B are independent if $\mathbb{P}(A|B) = \mathbb{P}(A)$.
2. Any set $A \subset \Omega$ and the empty set $\varnothing$ are independent, since $\mathbb{P}(A)\mathbb{P}(\varnothing) = 0 = \mathbb{P}(A \cap \varnothing)$.

The notion of independence can be generalized to any finite collection of sets.

Definition 2.3 Consider a collection of sets $\mathcal{A} = \{A_1, A_2, \ldots, A_n\}$. The collection is said to be

1. ***pairwise independent*** if, for any $i \neq j$,
$$\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j);$$
2. ***mutually independent*** if, for any sub-collection $\{A_{i_1}, \ldots, A_{i_k}\}$ of $\mathcal{A}$, where $k \leq n$,
$$\boxed{\mathbb{P}\left(\bigcap_{j=1}^{k} A_{i_j}\right) = \prod_{j=1}^{k} \mathbb{P}(A_{i_j}).} \tag{1}$$

Mutual independence implies pairwise independence, since any sub-collection consisting of two sets ($k = 2$) in $\mathcal{A}$ satisfies (1). However, the converse is in general NOT true.

The following formulas on conditional probabilities are useful:

1. (Law of Total Probability) For any $A \subset \Omega$, and any exhaustive and mutually exclusive events $\{B_i\}_{i=1}^n$ with $\mathbb{P}(B_i) > 0$ for $i = 1, 2, \ldots, n$, we have

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A|B_n)\mathbb{P}(B_n). \tag{2}$$

2. (Bayes Formula) For any $A, B \subset \Omega$ with $\mathbb{P}(A), \mathbb{P}(B) > 0$,

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}. \tag{3}$$

Equation (2) can be derived by using the law of total probability on p.1, and the definition of conditional probability; the Bayes formula (3) is an immediate consequence of the definition of conditional probability. In particular, the Bayes formula is fundamental in deducing the *posterior probability* in statistics and credibility theory.

Example 2.1 An insurer classifies drivers into three risk classes, Class 1 (low risk), Class 2 (medium risk), and Class 3 (H). The probability that a driver will be involved in a car accident in one year from each of the three risk classes are respectively 0.005 for Class 1, 0.01 for Class 2, and 0.2 from Class 3. Suppose that the insurer has a portfolio consisting of 90% of policyholders from Class 1, 7% from Class 2, and 3% from Class 3. It is known that a policy had been involved in a car accident last year. Find the probability that the policyholder is from Class 2.

Solution:

Let A be the event of having a policy had been involved in an accident last year. We are asked to compute $\mathbb{P}(\text{Class } 2|A)$. By the information, we can draw the following table:

Class i	1	2	3
$\mathbb{P}(A\|\text{Class } i)$	0.005	0.01	0.2
$\mathbb{P}(\text{Class } i)$	0.9	0.07	0.03
$\mathbb{P}(A\|\text{Class } i)\mathbb{P}(\text{Class } i)$	0.0045	0.0007	0.006

By Equation (2), we have

$$\mathbb{P}(A) = \sum_{i=1}^{3} \mathbb{P}(A|\text{Class } i)\mathbb{P}(\text{Class } i) = 0.0045 + 0.0007 + 0.006 = 0.0112.$$

By Bayes formula (Equation (3)), we have

$$\mathbb{P}(\text{Class } 2|A) = \frac{\mathbb{P}(A|\text{Class } 2)\mathbb{P}(\text{Class } 2)}{\mathbb{P}(A)} = \frac{0.0007}{0.0112} = 0.0625.$$

3 Random Variables and Distribution Functions

Random variables are mappings from the set of possible outcomes of a random event to a real number for numerical calculations.

Definition 3.1 A function $X : \Omega \mapsto \mathbb{R}$ is called a *random variable*.

In this course, we will study random variables to model frequency, severity, and aggregate loss of short-term insurances:

- **Frequency Models:** number of claims received within a planning horizon;
- **Severity Models:** the size/loss of an individual claim;
- **Aggregate Loss Models:** aggregating the losses over all claims received.

Example 3.1 Consider the random experiment of flipping a coin twice, with H denotes getting a head, and T denotes getting a tail. Then, $\Omega = \{\omega_1 = HH, \omega_2 = HT, \omega_3 = TH, \omega_4 = TT\}$. Consider the following two examples of random variables:

1. Let X be the random variable of number of heads obtained. Then $X(\omega_1) = 2$, $X(\omega_2) = X(\omega_3) = 1$, and $X(\omega_4) = 0$.
2. Let Y be the random variable such that $Y = 1$ if there is at least one head in the two flips, and $Y = 0$ otherwise. Then $Y(\omega_1) = Y(\omega_2) = Y(\omega_3) = 1$, and $Y(\omega_4) = 0$. If we let $A \subset \Omega$ be the set containing the outcomes with at least one head. Then $A = \{\omega_1, \omega_2, \omega_3\}$, and we can write Y as

$$Y(\omega) = \mathbb{1}_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A, \\ 0, & \text{otherwise.} \end{cases}.$$

   Y is also called an *indicator random variable*. This type of variables is very useful in probability theory.

Given a random variable X, we want to ask questions about how likely X would fall within a certain range $A \subset \mathbb{R}$. For example, what is the probability that a claim will be received from an insurance policy? What is the probability that the claim size would exceed 10 thousands? Knowing the *distributions* of the frequency and severity variables allows insurers perform pricing and reserving adequately. In particular, the distribution of a random variable X can be characterized by the following functions.

Definition 3.2 The ***cumulative distribution function*** (cdf) F_X of a random variable X, and the *survival function* S_X, are defined respectively by

$$\boxed{F_X(x) := \mathbb{P}(X \leq x) \quad \text{and} \quad S_X(x) := \mathbb{P}(X > x) = 1 - F_X(x).}$$

Why do cdf and survival function matter?

- **Characterization**: the entire distribution of X can be recovered from F_X.
- **Risk management and reserving**:
  - Quantile reserve: the fund we should keep such that the probability of being able to cover a loss X is high (say 99%);
  - the *tail* (*heavy* vs. *light*) of the distribution tells us the likelihood of extreme events.
- **Transformations** (Chapter 2)
  - The distribution of $Y = g(X)$ could be derived from the cdf F_X of X;
  - X can be simulated by the inverse transform sampling $F_X^{-1}(U)$, where U is uniformly distributed on $[0, 1]$.

The following characterizes a cdf of a random variable, whose proof is out of the scope of our course.

> **Theorem 3.1** A function $F(x)$ is a cdf if and only if the following conditions hold:
> 1. F is non-decreasing: $F(a) \leq F(b)$ if $a \leq b$;
> 2. $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to +\infty} F(x) = 1$;
> 3. F is right-continuous: $\lim_{x \to x_0^+} F(x) = F(x_0)$.

Depending on the support of X, we mainly classify random variables into the following classes in this course:

1. **Discrete Random Variables**
   - X is said to be a discrete random variable if its ***support***[1] $\text{Supp}(X)$ is countable, e.g., $\text{Supp}(X) = \mathbb{N}, \{1, 2, 3\}$ etc.
   - The distribution of a discrete random variable X can be expressed by its ***probability mass function (pmf)***:

     $$\boxed{p_X(x) = p_x = \mathbb{P}(X = x),\ \ x \in \text{Supp}(X),}$$

     which satisfies

     (a) $0 \leq p_x \leq 1$ for all $x \in \text{Supp}(X)$;

     (b) $\sum_{x \in \text{Supp}(X)} p_x = 1$.
   - For $A \subset \mathbb{R}$, the probability that $X \in A$ is given by

     $$\boxed{\mathbb{P}(X \in A) = \sum_{x \in A} p_x.}$$

[1] Roughly speaking, the support of a distribution is the set $\text{Supp}(X)$ such that its pmf/pdf is non-zero.

- Some common discrete distributions: Bernoulli, Binomial, Negative Binomial, etc. We will look at them in details when we study frequency models.

2. **Continuous Random Variables**

   - The support $\text{Supp}(X)$ of a continuous random variable X is uncountable (continuum), e.g., $\text{Supp}(X) = \mathbb{R}, [0, 1]$ etc.
   - The distribution of a continuous random variable X can be expressed by its ***probability density function (pdf)***:

   $$\boxed{f_X(x) := \frac{d}{dx} F_X(x) = F_X'(x),}$$

   where F_X is the cdf of X, such that

   (a) $f_X(x) \geq 0$ for all $x \in \text{Supp}(X)$;

   (b) $\int_{\text{Supp}(X)} f(x)dx = 1$.

   - Notice that $f_X(x)$ does NOT indicate the probability of $X = x$. It only gives the likelihood of X falling in a neighbourhood of x. Looking it in an infinitesimal way, given a small number dx,

   $$f_X(x)dx \approx \mathbb{P}(x < X < x + dx).$$

   Since $f_X(x)$ itself is NOT a probability, we do not require $f_X(x) \leq 1$.

   - For $A \subset \mathbb{R}$, the probability that $X \in A$ is given by

   $$\boxed{\mathbb{P}(X \in A) = \int_A f_X(x)dx.}$$

   For this reason, $\mathbb{P}(X = a) = 0$ for any $a \in \mathbb{R}$, since $\int_{\{a\}} f(x)dx = \int_a^a f_X(x)dx = 0$.

   - Some common continuous distributions: Uniform, Gamma, Gaussian, Pareto, Exponential, Beta, etc. We will look at them in details when we study severity models.

3. **Mixed/Compound Random Variables**

   - A mixed random variable consists of a discrete part, and a continuous part. It is more commonly seen for aggregate loss models.
   - It is not uncommon that no claims would be received for an insurance policy within a short period of time, and if it does receive claims, the claim size can be modelled by continuous distribution. Hence, the random variable of aggregate claim size S has a non-zero probability at 0, $\mathbb{P}(S = 0) > 0$; and the distribution can be expressed by a density function for loss $s > 0$.

- The Tweedie distribution is a class of distribution that includes mixed distributions, which have positive mass at 0.

Example 3.2 Consider a continuous random variable X with pdf

$$f_X(x) = \begin{cases} \dfrac{A}{x^{\alpha+1}}, & \text{if } x \geq B, \\ 0, & \text{if } x < B, \end{cases}$$

where $\alpha, A, B > 0$.

(a) Find the relationship between A and B such that f_X is a valid pdf.

(b) If $\alpha = B = 2$, find the probability $\mathbb{P}(3 < X < 5)$.

<u>*Solution:*</u>

(a) f_X is a valid pdf only if it integrates to 1. Hence,

$$1 = \int_B^\infty \frac{A}{x^{\alpha+1}} dx = -\frac{A}{\alpha x^\alpha}\bigg|_B^\infty = \frac{A}{\alpha B^\alpha}.$$

Hence, $A = \alpha B^\alpha$

(b) By (a), we know that $A = 2(2)^2 = 8$. Hence,

$$\mathbb{P}(3 < X < 5) = \int_3^5 \frac{8}{x^3} dx = -\frac{4}{x^2}\bigg|_3^5 = \frac{64}{225}.$$

Example 3.3 Find the cdf of the following random variables:

(a) A discrete random variable X with pmf

$$\mathbb{P}(X = x) = \begin{cases} 0.2, & \text{if } x = 0; \\ 0.4, & \text{if } x = 1; \\ 0.1, & \text{if } x = 3; \\ 0.3, & \text{if } x = 6. \end{cases}$$

(b) A mixed random variable Y with $\mathbb{P}(Y = 0) = 0.3$, $\mathbb{P}(Y = 10) = 0.2$, and for $y \in (0, 10)$, it admits the following density function:

$$f_Y(y) = \frac{3\sqrt{10y}}{400}, \quad 0 < y < 10.$$

Solution:

(a) The cdf of X is given by

$$F_X(x) = \begin{cases} 0, & \text{if } x < 0; \\ 0.2, & \text{if } 0 \le x < 1; \\ 0.2 + 0.4 = 0.6, & \text{if } 1 \le x < 3; \\ 0.2 + 0.4 + 0.1 = 0.7, & \text{if } 3 \le x < 6, \\ 0.2 + 0.4 + 0.1 + 0.3 = 1, & \text{if } x \ge 6. \end{cases}$$

(b) For $y \in (0, 10)$, we have

$$\int_0^y f_Y(z)dz = \int_0^y \frac{3\sqrt{10z}}{400} dz = \frac{\sqrt{10}y^{\frac{3}{2}}}{200}.$$

Hence, the cdf of Y is given by

$$F_Y(y) = \begin{cases} 0, & \text{if } y < 0; \\ 0.3 + \dfrac{\sqrt{10}y^{\frac{3}{2}}}{200}, & \text{if } 0 \le y < 10; \\ 0.3 + \dfrac{\sqrt{10} \times 10^{\frac{3}{2}}}{200} + 0.2 = 1, & \text{if } y \ge 10. \end{cases}$$

4 Distributional Quantities of Random Variables

This section is devoted to reviewing some basic distributional quantities of a random variable. In particular, we will be looking at the expected values, central moments, and percentiles. These values allow us to examine the central tendency and variability of a random variable.

4.1 Expected Values, Moments and Central Moments

The expected value of a random variable X gives the average value of X weighted by its distribution. Moments and central moments are expected values of functions of X.

1. **Expected Value:**
   - The expected value of X, denoted by μ or $\mathbb{E}[X]$, is defined as

$$\mu = \mathbb{E}[X] := \begin{cases} \displaystyle\int_{-\infty}^{\infty} x f_X(x)dx, & \text{if } X \text{ is continuous;} \\ \displaystyle\sum_x x p_X(x), & \text{if } X \text{ is discrete.} \end{cases}$$

- If $\mathbb{E}[X]$ exists as a finite number, we say that X is *integrable*.
- Expected value is a *linear operator*: for any $\alpha, \beta \in \mathbb{R}$, and any integrable random variables X, Y,

$$\boxed{\mathbb{E}[\alpha X + \beta Y] = \alpha \mathbb{E}[X] + \beta \mathbb{E}[Y].}$$

- In general, for any function $g(x)$,

$$\mathbb{E}[g(X)] := \begin{cases} \displaystyle\int_{-\infty}^{\infty} g(x) f_X(x)dx, & \text{if } X \text{ is continuous;} \\ \displaystyle\sum_x g(x) p_X(x), & \text{if } X \text{ is discrete.} \end{cases}$$

2. **Moments:**
   - For $k = 1, 2, \dots$, the k-th moment of X, denoted by μ'_k, is the expected value of X^k:

$$\boxed{\mu'_k := \mathbb{E}[X^k].}$$

   - By *Jensen's inequality*, one can show that X^{k+1} is integrable implies X^k is integrable.
3. **Central Moments:**
   - The k-th central moment of X, denoted by μ_k, is the expected value of $(X - \mu)^k$:

$$\boxed{\mu_k := \mathbb{E}[(X - \mu)^k].}$$

   - The second central moment of X is also known as *variance*:

$$\boxed{\mathrm{Var}[X] := \mu_2 = \mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2] - \mathbb{E}^2[X] = \mu'_2 - \mu^2.}$$

   For any $\alpha, \beta \in \mathbb{R}$, we have

$$\boxed{\mathrm{Var}[\alpha X + \beta] = \alpha^2 \mathrm{Var}[X].}$$

A few useful formulas concerning expected values are listed below:

1. For any set $A \subset \mathbb{R}$,
$$\mathbb{P}(X \in A) = \mathbb{E}[\mathbb{1}_{\{X \in A\}}], \tag{4}$$
where $\mathbb{1}_{\{x \in A\}}$ is the indicator function: $\mathbb{1}_{\{x \in A\}} = 1$ if $x \in A$, and 0 otherwise.
2. For any <u>positive</u> random variable X,
$$\mathbb{E}[X] = \int_0^\infty S_X(x)dx. \tag{5}$$

Proof.

1. For simplicity, we only consider X being a continuous random variable with density f_X. The first statement can be proven by the writing the probability in terms of an integral:
$$\mathbb{P}(X \in A) = \int_A f_X(x)dx = \int_{-\infty}^\infty \mathbb{1}_{\{x \in A\}} f_X(x)dx = \mathbb{E}[\mathbb{1}_{\{X \in A\}}].$$
2. Using Equation (4), we know that
$$S_X(x) = \mathbb{P}(X > x) = \mathbb{E}[\mathbb{1}_{\{X > x\}}].$$
Hence, by Fubini's theorem,
$$\int_0^\infty S_X(x)dx = \int_0^\infty \mathbb{E}[\mathbb{1}_{\{X > x\}}]dx = \mathbb{E}\left[\int_0^\infty \mathbb{1}_{\{X > x\}}dx\right] = \mathbb{E}\left[\int_0^X dx\right] = \mathbb{E}[X].$$

□

Equation (5) can be further generalized into higher order moments, whose proof is left as an exercise:

Theorem 4.1 Let $\alpha > 0$ and X be a positive random variable. If X^α is integrable,
$$\boxed{\mathbb{E}[X^\alpha] = \int_0^\infty \alpha x^{\alpha-1} S_X(x)dx.}$$

Example 4.1 Compute the expected value and the variance for the random variables X, Y in Example 3.3.

Solution:

(a) $\mathbb{E}[X] = 0 \times 0.2 + 1 \times 0.4 + 3 \times 0.1 + 6 \times 0.3 = 2.5$,
$\mathbb{E}[X^2] = 0^2 \times 0.2 + 1^2 \times 0.4 + 3^2 \times 0.1 + 6^2 \times 0.3 = 12.1$.
Hence, $\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}^2[X] = 12.1 - 2.5^2 = 5.85$.

(b) For the random variable Y,

$$\begin{aligned}
\mathbb{E}[Y] &= 0 \times \mathbb{P}(Y = 0) + \int_0^{10} y f_Y(y) dy + 10 \times \mathbb{P}(Y = 10) \\
&= \int_0^{10} \frac{3\sqrt{10} y^{\frac{3}{2}}}{400} dy + 10 \times 0.2 \\
&= 3 + 2 = 5, \\
\mathbb{E}[Y^2] &= 0^2 \times \mathbb{P}(Y = 0) + \int_0^{10} y^2 f_Y(y) dy + 10^2 \times \mathbb{P}(Y = 10) \\
&= \int_0^{10} \frac{3\sqrt{10} y^{\frac{7}{2}}}{400} dy + 10^2 \times 0.2 \\
&= \frac{150}{7} + 20 = \frac{290}{7}.
\end{aligned}$$

Therefore, $\text{Var}[X] = \frac{290}{7} - 5^2 = \frac{115}{7}$.

We end this subsection by listing some useful distributional quantities:

Quantity	Definition	Symbol
Variance	μ_2	σ^2
Standard deviation	$\sqrt{\mu}$	σ
Coefficient of variation (C.V.)	σ/μ	/
Skewness	μ_3/σ^3	γ_1
Kurtosis	μ_4/σ^4	γ_2

Table 1: Some other distributional quantities

A few remarks on the quantities introduced in Table 1:

- Skewness is a measure of *asymmetry*: a symmetric distribution has a zero skewness, and a distribution with a positive (resp. negative) skewness has a long right (resp. left) tail.
- Kurtosis of a distribution is a measure of *heaviness of its tail (tailedness)*. A random variable with a high (resp. low) kurtosis tends to have heavy (light) tails, or outliers.
- C.V., skewness and kurtosis are unit-less and scale-invariant, i.e., the kurtosis and

skewness of X and cX are the same for any $c > 0$.

4.2 Percentiles

Percentiles of a distribution is the inverse of its cdf.

Definition 4.1 For $p \in (0,1)$, the $100p$-th percentile of a random variable X is defined as

$$\boxed{x_p := \inf\{x : F_X(x) \geq p\}.}$$

A few remarks on percentiles:

- The $100\%p$-th percentile is also called the *p-quantile*.
- We can check whether an unknown distribution is given by a candidate distribution by computing the percentiles. This can be done by comparing the empirical percentiles of the unknown distribution, and the theoretical percentiles of the known candidate via a *qq-plot*.
- When F_X is strictly increasing and continuous, $x_p = F_X^{-1}(p)$. Otherwise, the inverse of F_X may not exist. In general, we (only) have $F_X(x_p) \geq p$.
- Percentile is a *risk measure*, which is equivalent to the *Value-at-Risk*. We shall come back to it when we study risk measures.

Example 4.2 Find the 30-th, 60-th, and the 90-th percentile of the random variables X and Y in Example 3.3.

Solution:

(a) $F_X(x) < 0.3$ for $0 \leq x < 1$, and $F_X(1) = 0.6 > 0.3$. Hence, $x_{0.3} = 1$. This also implies that $x_{0.6} = 1$. Finally, $F_X(x) = 0.7 < 1$ for $x < 6$, and $F_X(6) = 1 > 0.9$. Hence, $x_{0.9} = 6$.

(b) $F_Y(y) = 0$ for $y < 0$, and $F_Y(0) = 0.3$. Hence, $y_{0.3} = 0$. To find the 60-th percentile, by setting

$$0.6 = F_Y(y_{0.6}) = 0.3 + \frac{\sqrt{10}y_{0.6}^{\frac{3}{2}}}{200}$$

$$\Rightarrow y_{0.6} = \left(\frac{200(0.6-0.3)}{\sqrt{10}}\right)^{\frac{2}{3}} = \sqrt[3]{360} = 7.1138 \in (0,10).$$

Finally, $\lim_{y\to 10^-} F_Y(y) = 0.8 < 0.9$, and $F_Y(10) = 1 > 0.9$. Therefore, $y_{0.9} = 10$.

5 Generating Functions

In this section, we introduce two generating functions – ***the moment generating function* (mgf)**, and the ***probability generating function* (pgf)**. These functions are very useful tools in this course:

1. They characterize the distribution of a function: two random variables with the same generating functions must have the same distribution.
2. They can be used to compute moments/probabilities.
3. They allow simple characterization of *compound distributions* when we study frequency and aggregate risk models.

5.1 Moment Generating Functions

Definition 5.1 The ***moment generating function* (mgf)** of a random variable is defined by

$$M_X(t) := \mathbb{E}[e^{tX}], \ t \in \mathbb{R}.$$

Remark 5.1. If X is a continuous random variable, the mgf of X is equivalent to the *two-sided Laplace transform* of its density function f_X.

As suggested by its name, the mgf of X can be used to compute moments by taking its derivatives.

Theorem 5.2 For $k \in \mathbb{N}$ and a random variable X,

$$\boxed{\mu'_k = \mathbb{E}[X^k] = M_X^{(k)}(0) = \frac{d^k}{dt^k} M_X(t)\bigg|_{t=0}.}$$

Proof. Under some regularity conditions[2], we can switch the order of differentiation and expectation. Hence,

$$M_X^{(k)}(t) = \frac{d^k}{dt^k} M_X(t) = \frac{d^k}{dt^k} \mathbb{E}[e^{tX}] = \mathbb{E}\left[\frac{d^k}{dt^k} e^{tX}\right] = \mathbb{E}[X^k e^{tX}].$$

Therefore,

$$M_X^{(k)}(0) = \mathbb{E}[X^k e^{0\times X}] = \mathbb{E}[X^k].$$

□

[2] See the Leibniz integral rule, for example.

5.2 Probability Generating Functions

If X is a discrete random variable, the pgf could be more useful than a mgf. Indeed, the pgf can be obtained from the mgf by a change of variable.

Definition 5.2 The ***probability generating function*** **(pgf)** of a random variable is defined by

$$P_X(t) := \mathbb{E}[t^X].$$

Notice that the mgf and the pgf are related by $M_X(t) = P_X(e^t)$. If X is a discrete random variable taking values in $\mathbb{N}_0 = \{0, 1, 2, \dots\}$, we can write its pgf as

$$\boxed{P_X(t) = \mathbb{E}[t^X] = \sum_{k=0}^{\infty} t^k p_k = p_0 + tp_1 + t^2 p_2 + \cdots .}$$

Using this representation, we see that

1. $P_X(0) := \lim_{t\to 0^+} P_X(t) = p_0$;
2. $P_X(1) = 1$;
3. $P_X^{(k)}(t) = k!p_k + \sum_{j=1}^{\infty} \frac{(k+j)!}{j!} t^j p_{k+j}$.

Using the third property, we see that the pgf of X can be used to generate its pmf:

Theorem 5.3 Consider a discrete random variable X taking values in $\mathbb{N}_0$. For any $k \in \mathbb{N}_0$,

$$\boxed{p_k = \frac{P_X^{(k)}(0)}{k!}}.$$

Example 5.1 Find the pgf of the random variable X in Example 3.3.
<u>*Solution:*</u>
The pgf of X is given by

$$P_X(t) = p_0 + p_1 t^1 + p_3 t^3 + p_6 t^6 = 0.2 + 0.4t + 0.1t^3 + 0.3t^6, \ t \in \mathbb{R}.$$

Example 5.2 Let X be a discrete random variable taking values in $\mathbb{N}_0 = \{0, 1, 2, \dots\}$. Let $\lambda > 0$ be a parameter. Suppose that the pgf of X is given by

$$P_X(t) = e^{\lambda(t-1)}, t \in \mathbb{R}.$$

(a) Calculate $p_X(5) = \mathbb{P}(X = 5)$.
(b) Calculate $\mathrm{Var}[X]$.

Solution:

(a) It is easy to see that
$$P_X^{(k)}(t) = \lambda^k e^{\lambda(t-1)}.$$
Hence,
$$p_X(5) = \frac{\lambda^5 e^{-\lambda}}{5!} = \frac{\lambda^5 e^{-\lambda}}{120}.$$

(b) We compute $\mathrm{Var}[X]$ by the mfg of X. Note that
$$M_X(t) = P_X(e^t) = e^{\lambda(e^t-1)}.$$
Using this, we have
$$M_X'(t) = \lambda e^t e^{\lambda(e^t-1)} \Rightarrow M_X'(0) = \mathbb{E}[X] = \lambda,$$
$$M_X''(t) = \lambda e^t e^{\lambda(e^t-1)} + \lambda^2 e^{2t} e^{\lambda(e^t-1)} \Rightarrow M_X''(0) = \mathbb{E}[X^2] = \lambda + \lambda^2.$$
Therefore,
$$\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}^2[X] = \lambda + \lambda^2 - \lambda^2 = \lambda.$$

6 Joint Distributions

In this section, we shall discuss the distribution of a number of random variables collectively.

6.1 Bivariate Distributions

Let X and Y be two random variables. The *joint distribution* of X and Y is the probability distribution of all possible pairs of outcomes (X, Y). In the sequel, we let $\mathrm{Supp}(X)$ and $\mathrm{Supp}(Y)$ be the support of X and Y, respectively,

- If X and Y are both discrete random variables, we can describe the distribution of (X, Y) by the *joint probability mass function*:
$$\boxed{p_{X,Y}(j,k) = p_{j,k} = \mathbb{P}(X = j, Y = k),}$$
such that
  1. $0 \le p_{j,k} \le 1$;
  2. $\sum_j \sum_k p_{j,k} = 1$.

For any $A \subset \mathbb{R} \times \mathbb{R}$, the probability that the pair (X, Y) is in A is given by

$$\boxed{\mathbb{P}((X,Y) \in A) = \sum_{(j,k)\in A} p_{j,k}.}$$

- If X and Y are both continuous random variables, we can describe the distribution of (X, Y) by the *joint probability density function*:

$$\boxed{f_{X,Y}(x, y),}$$

  such that

  1. $f_{X,Y}(x, y) \geq 0$ for any $x, y \in \mathbb{R}$;
  2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y)dxdy = 1.$

  For any $A \subset \mathbb{R} \times \mathbb{R}$, the probability that the pair (X, Y) is in A is given by

$$\boxed{\mathbb{P}((X,Y) \in A) = \int\int_{(x,y)\in A} f_{X,Y}(x, y)dxdy}$$

The standalone distribution of X (or Y), i.e., the *marginal distribution*, can be obtained from their joint distribution:

- If X and Y are both discrete random variable with joint pmf, the *marginal pmfs* can be obtained as follows:

$$\boxed{p_X(x) = \sum_{y\in \text{Supp}(Y)} p_{X,Y}(x, y),\ \ p_Y(y) = \sum_{x\in \text{Supp}(X)} p_{X,Y}(x, y).}$$

- If X and Y are both continuous random variable with joint pdf $f_{X,Y}$, we can obtain the *marginal pdfs* as follows:

$$\boxed{f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy,\ \ f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dx.}$$

Like univariate random variables, we can define the joint cdf of the pair (X, Y) which characterizes the joint distribution.

Definition 6.1 The joint cdf of (X, Y) is defined as

$$\boxed{F_{X,Y}(x, y) := \mathbb{P}(X \leq x, Y \leq y).}$$

Remark 6.1. For multivariate random variables, $\mathbb{P}(X > x, Y > y) \neq 1 - F_{X,Y}(x, y)$. By inclusion-exclusion, the correct relationship is given by

$$\mathbb{P}(X > x, Y > y) = 1 - (F_X(x) + F_Y(y) - F_{X,Y}(x, y)).$$

Given a function $g : \mathbb{R}^2 \to \mathbb{R}$, we can compute the expected value of $g(X, Y)$ as follows:

$$\boxed{\mathbb{E}[g(X,Y)] = \begin{cases} \sum_x \sum_y g(x,y) p_{X,Y}(x,y), & \text{if } X, Y \text{ are discrete;} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y) dx dy, & \text{if } X, Y \text{ are continuous.} \end{cases}}$$

One way to measure the dependence of X and Y is to compute their *covariance*, and the *correlation coefficient*:

Definition 6.2

1. The **covariance** of X and Y is defined as

$$\boxed{\mathrm{Cov}(X, Y) := \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}[XY] - \mu_X \mu_Y,}$$

   where $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$.

2. The (Pearson) **correlation coefficient** of X and Y is defined as

$$\boxed{\rho_{X,Y} := \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y},}$$

   where σ_X and σ_Y are the standard deviations of X and Y, respectively.

Remark 6.2.

1. The covariance and the correlation coefficient measure the *linear dependence* of X and Y.
2. By the Cauchy–Schwarz inequality, we must have $-1 \leq \rho_{X,Y} \leq 1$;
3. If $\rho_{X,Y} = 1$ (resp. -1), then there exists $a > 0$ (resp. $a < 0$), and $b \in \mathbb{R}$, such that $Y = aX + b$.

6.2 Multivariate Distributions

The notion of joint distributions can be generalized to more than two random variables. Let $\{X_i\}_{i=1}^n = \{X_1, X_2, \ldots, X_n\}$ be a collection of random variables (a.k.a. *random vector*). Depending on the support of the $\{X_i\}_{i=1}^n$, the joint distribution of the random vector can be discussed as follows:

- If the distribution of each X_i is discrete, we can characterize the joint distribution by the *joint pmf*:

$$\boxed{p_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \mathbb{P}(X_1 = x_1,\ldots,X_n = x_n).}$$

- If the distribution of each X_i is continuous, we can characterize the joint distribution by the *joint pdf*:

$$\boxed{f_{X_1,\ldots,X_n}(x_1,\ldots,x_n).}$$

- In either case, we can define the *joint cdf* as:

$$\boxed{F_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \mathbb{P}(X_1 \le x_1,\ldots,X_n \le x_n).}$$

6.3 Conditional Distributions

If X and Y are dependent of each other, the additional knowledge of one variable will change the distribution of the other. This can be described by the *conditional distribution*, which can be defined under a similar spirit as conditional probabilities in Definition 2.1.

The ***conditional distribution of*** X ***given*** Y, denoted by $X|Y$, can be characterized as follows.

1. If X and Y are discrete random variables:
   - The ***conditional pmf*** of X given Y is defined as

$$\boxed{p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)} = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)}.}$$

2. If X and Y are continuous random variables:
   - the ***conditional pmf*** of X given Y is defined as

$$\boxed{f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}.}$$

Given that $Y = y$, we write $X|Y = y$ as the conditional distribution of X given $Y = y$. It is a distribution of X!

Like independence of two events, if the knowledge of Y does not change the distribution of X, we say that X and Y are *independent*.

Definition 6.3 The random variables X and Y are said to be ***independent***, denoted by $X \perp\!\!\!\perp Y$, if

$$\boxed{F_{X,Y}(x,y) = F_X(x)F_Y(y).}$$

If X, Y are both discrete or both continuous, we can also define independence of X and Y as follows:

- If X and Y are both discrete random variables, then $X \perp\!\!\!\perp Y$ iff

$$\boxed{p_{X,Y}(x,y) = p_X(x)p_Y(y),\ \forall x \in \text{Supp}(X), y \in \text{Supp}(Y).}$$

  If $p_Y(y) > 0$, this relation can also be written as $p_{X|Y}(x|y) = p_X(x)$, i.e., the conditional pmf of $X|Y = y$ is just p_X, which implies the knowledge of Y does not change the distribution of X.

- If X and Y are both continuous random variables, then $X \perp\!\!\!\perp Y$ iff

$$\boxed{f_{X,Y}(x,y) = f_X(x)f_Y(y),\ \forall x \in \text{Supp}(X), y \in \text{Supp}(Y).}$$

  If $f_Y(y) > 0$, this relation can also be written as $f_{X|Y}(x|y) = f_X(x)$.

Proposition 6.3 Let X, Y be two random variables, and f, g be two functions such that $g(X)$ and $h(Y)$ are integrable. If $X \perp\!\!\!\perp Y$,

$$\boxed{\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)].}$$

Proof. Since $X \perp\!\!\!\perp Y$, we have $f_{X,Y}(x,y) = f_X(x)f_Y(y)$. Hence,

$$\begin{aligned}
\mathbb{E}[g(X)h(Y)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f_X(x)f_Y(y)dxdy \\
&= \left(\int_{-\infty}^{\infty} g(x)f_X(x)dx\right)\left(\int_{-\infty}^{\infty} h(y)f_Y(y)dy\right) \\
&= \mathbb{E}[g(X)]\mathbb{E}[h(Y)].
\end{aligned}$$

□

As an immediate consequence of Proposition 6.3, if $X \perp\!\!\!\perp Y$, $\text{Cov}(X,Y) = 0$. However, the converse is not true: $\text{Cov}(X,Y) = 0$ does NOT imply $X \perp\!\!\!\perp Y$; see the following example.

Example 6.1 Let X be a discrete random variable with pmf $p_X(-1) = 0.4$, $p_X(0) = 0.2$, and $p_X(1) = 0.4$. Define $Y := \mathbb{1}_{\{X=0\}}$.

(a) Find the pmf of Y, and the joint pmf of (X, Y). Is $X \perp\!\!\!\perp Y$?

(b) Compute $\text{Cov}(X, Y)$.

<u>*Solution:*</u>

(a) Y is a discrete variable with only two values, 1 or 0. $Y = 1$ if $X = 0$ with probability $p_X(0) = 0.2$; $Y = 0$ if $X \neq 0$ with probability $1 - p_X(0) = 0.8$. Hence, the pmf of Y is $p_Y(0) = 0.2$, and $p_Y(1) = 0.8$.
To find the joint pmf of (X, Y), notice that $X = 0$ implies $Y = 1$, and $X \neq 0$ implies $Y = 0$. The only possible outcomes are thus $(X, Y) = (0, 1), (-1, 0)$ and $(1, 0)$. The joint pmf is thus

$$p_{X,Y}(0, 1) = p_X(0) = 0.2,\ p_{X,Y}(-1, 0) = p_X(-1) = 0.4,\ p_{X,Y}(1, 0) = p_X(1) = 0.4.$$

Notice that

$$p_X(1)p_Y(1) = 0.4 \times 0.8 = 0.32 \neq 0 = p_{X,Y}(1, 1).$$

Hence, X and Y are not independent.

(b) Notice that $\mathbb{E}[X] = 0$. Also, $\mathbb{P}(XY = 0) = 1$, which implies $\mathbb{E}[XY] = 0$. Hence, $\text{Cov}(X, Y) = 0$.

The notion of independence can be extended to a random vector $\{X_i\}_{i=1}^n$.

Definition 6.4 A collection of random variables $\{X_i\}_{i=1}^n$ is said to be ***mutually independent*** if

$$\boxed{F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^n F_{X_i}(x_i).}$$

Remark 6.4.

1. If each X_i is discrete, $\{X_i\}_{i=1}^n$ is mutually independent if its pmf satisfies the following:

$$\boxed{p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^n p_{X_i}(x_i).}$$

2. If each X_i is continuous, $\{X_i\}_{i=1}^n$ is mutually independent if its pdf satisfies the following:

$$\boxed{f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^n f_{X_i}(x_i).}$$

3. The notion of mutual independence should not be confused with ***pairwise independence***. The collection $\{X_i\}_{i=1}^n$ is *pairwise independent* if $X_i \perp\!\!\!\perp X_j$ for any $i \neq j$. In general, mutual independence implies pairwise independence, but NOT vice verse.

4. If $\{X_i\}_{i=1}^n$ is mutually independent, and each X_i has the same distribution, we say that the collection is ***independent and identically distributed (i.i.d.)***.

7 Conditional Expectations

In this last section of Chapter 1, we review the notion of conditional expectations, which is fundamental in studying mixing, collective risk model, and credibility theory. We will define conditional expectations depending on whether the given knowledge is an event or a random variable.

7.1 Conditional Expectation Given an Event

Definition 7.1 Let X be a random variable, and $A \subset \Omega$ be an event with $\mathbb{P}(A) > 0$. The ***conditional expectation*** of X given A is defined as

$$\mathbb{E}[X|A] = \frac{\mathbb{E}[X\mathbb{1}_A]}{\mathbb{P}(A)} = \begin{cases} \dfrac{1}{\mathbb{P}(A)} \displaystyle\sum_{x \in A} x\mathbb{P}(X = x), & \text{if } X \text{ is discrete;} \\ \dfrac{1}{\mathbb{P}(A)} \displaystyle\int_A x f_X(x)dx, & \text{if } X \text{ is continuous.} \end{cases}$$

From Definition 7.1, we see that $\mathbb{E}[X|A]$ is a *real number*. Since we know that A has occurred, when computing the expected value of X, we are only counting those possible values x which overlaps with A, i.e., $\mathbb{E}[X\mathbb{1}_A]$ in the numerator. The denominator $\mathbb{P}(A)$ is the numeraire which accounts for the fact that we are confining ourselves to the given set A.

Example 7.1 Suppose that X is a continuous random variable with pdf

$$f_X(x) = \lambda e^{-\lambda x}, \ x > 0,$$

where $\lambda > 0$ is a constant. For any constant $d > 0$, compute $\mathbb{E}[X|X > d]$.
Solution:
First, we need to compute $\mathbb{P}(X > d)$:

$$\mathbb{P}(X > d) = \int_d^\infty \lambda e^{-\lambda x} dx = e^{-\lambda d}.$$

On the other hand,

$$\mathbb{E}[X\mathbb{1}_{\{X>d\}}] = \int_d^\infty \lambda x e^{-\lambda x} dx = e^{-\lambda d}\left(d + \frac{1}{\lambda}\right).$$

Therefore,

$$\mathbb{E}[X|X>d] = \frac{\mathbb{E}[X\mathbb{1}_{\{X>d\}}]}{\mathbb{P}(X>d)} = \frac{e^{-\lambda d}\left(d+\frac{1}{\lambda}\right)}{e^{-\lambda d}} = d + \frac{1}{\lambda}.$$

Quite often, we are interested in the expected value of X given the value of another related random variable, that is, $\mathbb{E}[X|A]$, where $A = \{Y = y\}$. If X, Y are discrete, we have $\mathbb{P}(A) = \mathbb{P}(Y = y)$, and $\mathbb{E}[X\mathbb{1}_{\{Y=y\}}] = \sum_x x p_{X,Y}(x, y)$ (show that!). However, when Y is continuous, $\mathbb{P}(Y = y) = 0$. In that case, we define the conditional expectation using the conditional density function.

Definition 7.2 Let X and Y be random variables. The ***conditional expectation*** of X given $Y = y$ is defined as

$$\boxed{\mathbb{E}[X|Y=y] = \begin{cases} \displaystyle\sum_x x p_{X|Y}(x|y), & \text{if } X, Y \text{ are discrete;} \\ \displaystyle\int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx, & \text{if } X, Y \text{ are continuous.} \end{cases}}$$

Example 7.2 Consider the random variables X and Y in Example 6.1. Find $\mathbb{E}[X|Y = y]$ for $y = 0$ and 1.

Solution:

If $Y = 1$, we must have $X = 0$. Hence, $\mathbb{E}[X|Y = 1] = 0$. If $Y = 0$, we know that X has equal probability of being -1 and 1. Hence, $\mathbb{E}[X|Y = 0] = -1 \times 0.5 + 1 \times 0.5 = 0$.

We can also define the conditional variance of X given $Y = y$.

Definition 7.3 The **conditional variance** of X given $Y = y$ is defined as

$$\boxed{\text{Var}[X|Y=y] = \mathbb{E}\left[(X - \mathbb{E}[X|Y=y])^2|Y=y\right] = \mathbb{E}[X^2|Y=y] - \mathbb{E}^2[X|Y=y].}$$

When the event $Y = y$ is given, the expected value of X would become $\mathbb{E}[X|Y = y]$. The conditional variance is thus defined as the conditional expectation of $(X - \mathbb{E}[X|Y = y])^2$ (instead of $(X - \mathbb{E}[X])^2$) given $Y = y$.

7.2 Conditional Expectation Given a Random Variable

In Definition 7.2, we defined the conditional expectation $\mathbb{E}[X|Y=y]$, which evaluates the average value of X given that $Y=y$. As the possible outcome y of Y varies, the value $\mathbb{E}[X|Y=y]$ will also change accordingly. Hence, one can view the mapping $\mathbb{E}[X|Y=y]$ as a *function of* y. Let $h(y):=\mathbb{E}[X|Y=y]$. Then, the *random variable* $h(Y)$ is called the *conditional expectation* of X given Y:

Definition 7.4 Let X, Y be random variables. Define $h(y):=\mathbb{E}[X|Y=y]$. Then, the ***conditional expectation*** of X given Y is the random variable

$$\boxed{\mathbb{E}[X|Y]:=h(Y).}$$

Remark 7.1.

1. $\mathbb{E}[X|Y]=h(Y)$ is a **random variable**.
2. $\mathbb{E}[X|Y]=h(Y)$ is a **function of** Y, instead of X.
3. $\mathbb{E}[X|Y]$ is the *"best estimate"* of the random variable X based on Y in the $\mathcal{L}^2$-sense.
4. If $X \perp\!\!\!\perp Y$, $\mathbb{E}[X|Y=y]=\mathbb{E}[X]$. Hence $h(y)=\mathbb{E}[X]$ is a constant function, and $\mathbb{E}[X|Y]=\mathbb{E}[X]$.
5. If $X=g(Y)$, i.e., X is some transformation of Y, then $\mathbb{E}[X|Y]=g(Y)=X$. In other words, once we know Y, the best estimate of X is just X itself, since we can readily compute X by $X=g(Y)$.
6. In credibility theory, $\mathbb{E}[X|Y]$ is also called the *hypothetical mean*.

The following is a very important property concerning conditional expectations.

Theorem 7.2 (Law of iterated expectations/Tower property) For any random variables X and Y, we have

$$\boxed{\mathbb{E}\left[\mathbb{E}[X|Y]\right]=\mathbb{E}[X].}$$

Remark 7.3. In some textbooks, the law of iterated expectations is sometimes written as

$$\mathbb{E}[X]=\mathbb{E}_Y\left[\mathbb{E}_X[X|Y]\right].$$

Proof. Using the definition of conditional expectations and Fubini's theorem,

$$\begin{aligned}\mathbb{E}[\mathbb{E}[X|Y]]=\mathbb{E}[h(Y)]&=\int_{-\infty}^{\infty}h(y)f_Y(y)dy\\&=\int_{-\infty}^{\infty}\mathbb{E}[X|Y=y]f_Y(y)dy\end{aligned}$$

$$
\begin{aligned}
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx \right) f_Y(y) dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X|Y}(x|y) f_Y(y) dx dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X,Y}(x,y) dx dy \\
&= \int_{-\infty}^{\infty} x \left(\int_{-\infty}^{\infty} f_{X,Y}(x,y) dy \right) dx \\
&= \int_{-\infty}^{\infty} x f_X(x) dx = \mathbb{E}[X].
\end{aligned}
$$

□

Definition 7.5 The conditional variance of X given Y is defined as

$$
\boxed{\mathrm{Var}[X|Y] = \mathbb{E}\left[(X - \mathbb{E}[X|Y])^2|Y\right] = \mathbb{E}[X^2|Y] - \mathbb{E}^2[X|Y].}
$$

The (unconditional) variance of X can be decomposed into conditional ones as follows.

Theorem 7.4 (Law of total variance)

$$
\boxed{\mathrm{Var}[X] = \mathbb{E}\left[\mathrm{Var}[X|Y]\right] + \mathrm{Var}\left[\mathbb{E}[X|Y]\right].}
$$

$\mathrm{Var}[X]$ is the sum expected value of the conditional variance, and the variance of the conditional expectation. In credibility theory, the conditional variance is also called the *process variance.*

Proof. By the law of iterated expectations, we have

$$
\begin{aligned}
\mathrm{Var}[X] &= \mathbb{E}[(X-\mu)^2] \\
&= \mathbb{E}\left[\mathbb{E}\left[(X-\mu)^2|Y\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}[X^2|Y] - 2\mu\mathbb{E}[X|Y] + \mu^2\right] \\
&= \mathbb{E}[\mathbb{E}[X^2|Y] - \mathbb{E}^2[X|Y]] + \mathbb{E}\left[\mathbb{E}^2[X|Y] - 2\mu\mathbb{E}[X|Y] + \mu^2\right] \\
&= \mathbb{E}[\mathrm{Var}[X|Y]] + \mathbb{E}\left[\mathbb{E}^2[X|Y]\right] - 2\mu^2 + \mu^2 \\
&= \mathbb{E}[\mathrm{Var}[X|Y]] + \mathbb{E}[\mathbb{E}^2[X|Y]] - \mu^2 \\
&= \mathbb{E}[\mathrm{Var}[X|Y]] + \mathrm{Var}[\mathbb{E}[X|Y]],
\end{aligned}
$$

where the fifth line again follows from the law of iterated expectations, and the last line follows from the observation that

$$
\mathrm{Var}[\mathbb{E}[X|Y]] = \mathbb{E}[\mathbb{E}^2[X|Y]] - \mathbb{E}^2[\mathbb{E}[X|Y]] = \mathbb{E}[\mathbb{E}^2[X|Y]] - \mu^2.
$$

□

Example 7.3 Let X be the random variable of a loss (in thousands) from an insurance policy. Given the risk factor Θ, it is known that the pdf of X is given by

$$f_{X|\Theta}(x|\theta) = \frac{1}{2\theta} e^{-\frac{x}{2\theta}}, \ x > 0.$$

It is known that the risk factor Θ itself is a random variable with pmf

$$p_\Theta(1) = 0.6, \ p_\Theta(5) = 0.3, \ p_\Theta(10) = 0.1.$$

(a) Find the pmf of $\mathbb{E}[X|\Theta]$.
(b) Calculate $\mathbb{E}[X]$.
(c) Find the pmf of $\mathrm{Var}[X|\Theta]$.
(d) Calculate $\mathrm{Var}[X]$.

Solution:

(a) For $\theta = 1, 5$ and 10, we know that

$$\mathbb{E}[X|\Theta = \theta] = \int_0^\infty \frac{x}{2\theta} e^{-\frac{x}{2\theta}} dx = 2\theta.$$

Hence, the pmf of $\mathbb{E}[X|\Theta]$ is given by

$$\mathbb{P}\left(\mathbb{E}[X|\Theta] = y\right) = \begin{cases} 0.6, & \text{if } y = 2; \\ 0.3, & \text{if } y = 10; \\ 0.1, & \text{if } y = 20. \end{cases}$$

(b) By the law of iterated expectation,

$$\begin{aligned} \mathbb{E}[X] &= \mathbb{E}\left[\mathbb{E}[X|\Theta]\right] \\ &= \mathbb{E}[X|\Theta = 1]p_\Theta(1) + \mathbb{E}[X|\Theta = 5]p_\Theta(5) + \mathbb{E}[X|\Theta = 10]p_\Theta(10) \\ &= 2 \times 0.6 + 10 \times 0.3 + 20 \times 0.1 \\ &= 6.2. \end{aligned}$$

(c) For $\Theta = \theta$,

$$\mathrm{Var}[X|\Theta = \theta] = \mathbb{E}\left[X^2|\Theta = \theta\right] - \mathbb{E}^2[X|\Theta = \theta] = \mathbb{E}\left[X^2|\Theta = \theta\right] - (2\theta)^2,$$

where

$$\mathbb{E}[X^2|\Theta = \theta] = \int_0^\infty \frac{x^2}{2\theta} e^{-\frac{x}{2\theta}} dx = 8\theta^2.$$

Hence, $\mathrm{Var}[X|\Theta = \theta] = 8\theta^2 - 4\theta^2 = 4\theta^2$. The pmf of $\mathrm{Var}[X|\Theta]$ is thus

$$\mathbb{P}\left(\text{Var}[X|\Theta] = y\right) = \begin{cases} 0.6, & \text{if } y = 4; \\ 0.3, & \text{if } y = 100; \\ 0.1, & \text{if } y = 400. \end{cases}$$

(d) We compute $\text{Var}[X]$ by the conditional variance formula (Theorem 7.4). From (a), we can compute $\text{Var}[\mathbb{E}[X|\Theta]]$:

$$\begin{aligned} \text{Var}[\mathbb{E}[X|\Theta]] &= \mathbb{E}\left[(\mathbb{E}[X|\Theta])^2\right] - \left(\mathbb{E}\left[\mathbb{E}[X|\Theta]\right]\right)^2 \\ &= 2^2 \times 0.6 + 10^2 \times 0.3 + 20^2 \times 0.1 - \mathbb{E}^2[X] \\ &= 72.4 - 6.2^2 = 33.96. \end{aligned}$$

On the other hand, by (c),

$$\mathbb{E}[\text{Var}[X|\Theta]] = 4 \times 0.6 + 100 \times 0.3 + 400 \times 0.1 = 72.4.$$

Therefore, $\text{Var}[X] = \mathbb{E}[\text{Var}[X|\Theta]] + \text{Var}[\mathbb{E}[X|\Theta]] = 72.4 + 33.96 = 106.36$.